Two lessons from ensemble view on RNA structure

Teresa Przytycka

Algorithmic Methods
in Computational and Systems Biology
Research Group

NCBI/ NLM / NIH

Research topics of the Algorithmic Methods in Computational and Systems Biology group

- Network Approaches to cancer
- Inferring genotype-phenotype relations
- Gene regulation
- non-B-DNA structures
- RNA aptamers and their sequence/structure motifs

Two (and half) stories exploring ensemble view on RNA structure

- Network Approaches to cancer
- Inferring genotype-phenotype relations
  - impact of a SNV on mRNA structure
- Gene regulation
- non-B-DNA structures
- RNA aptamers and their sequence/structure motifs
  - Importance of the ensemble approach for delineating such motifs

Impact of mutations / single nucleotide variation (SNV) on RNA structure

Collaborators: C. Kimchi-Sarfaty FDA; M. Gottesman, NCI

A Silent Polymorphism in the MDR1 Gene Changes Substrate Specificity –
C. Kimchi-Sarfaty et al. Science 2006

Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. Chen, R., Davydov, E.V., Sirota, M., and Butte, A. J. PloS One 2010

Impact of mutations / single nucleotide variation (SNV) on RNA structure

Collaborators: C. Kimchi-Sarfaty FDA; M. Gottesman, NCI

A Silent Polymorphism in the MDR1 Gene Changes Substrate Specificity –
C. Kimchi-Sarfaty et al. Science 2006

Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. Chen, R., Davydov, E.V., Sirota, M., and Butte, A. J. PloS One 2010

Codon Usage?

Codon usage is proposed to bw optimized for variety of reasons e.g. avoiding frameshifting errors

Hoang, Koonin, Lipman, Przytycka, NAR 2008

Impact of mutations / single nucleotide variation (SNV) on RNA structure

Collaborators: C. Kimchi-Sarfaty FDA; M. Gottesman, NCI

A Silent Polymorphism in the MDR1 Gene Changes Substrate Specificity –
C. Kimchi-Sarfaty et al. Science 2006

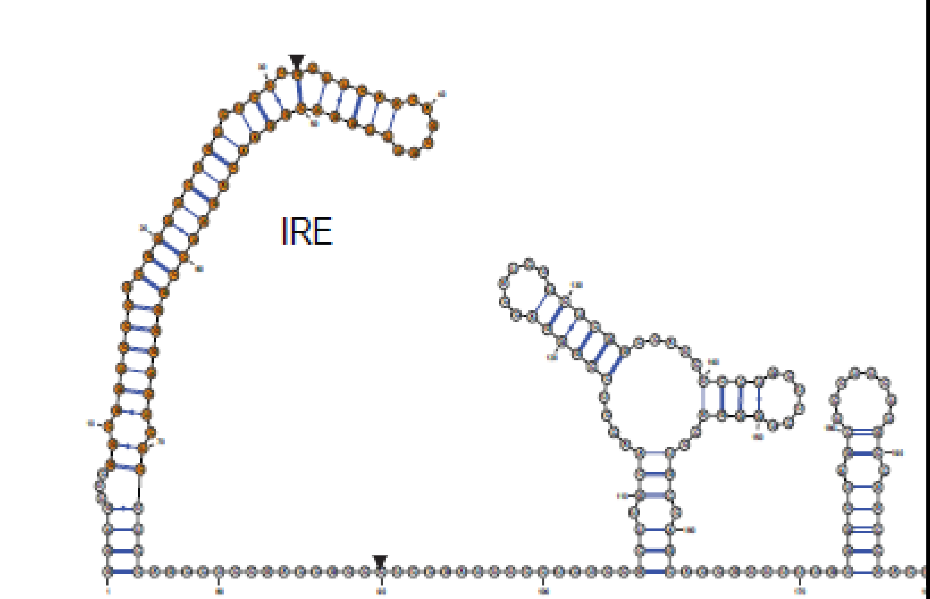
Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. Chen, R., Davydov, E.V., Sirota, M., and Butte, A. J. PloS One 2010

Possible results of SNP-induced mRNA structural changes

- changes in translation dynamics leading to altered folding kinetics and potentially protein misfolding
- impact on splicing
- 5'UTR structure has impact on gene expression
- Changes in other structurally important elements

Example

FTL light subunit of the ferritin protein



(a) Predicted Secondary Structure

The mutations that cause hyperferritinemia-cataract syndrome are found in a segment of the gene called the iron responsive element (IRE)

Challenge: how to measure structural impact of an SNV

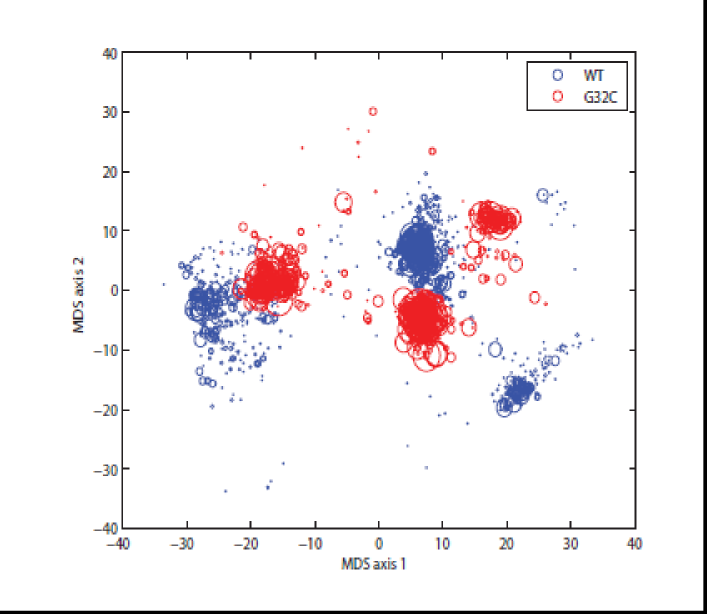
- Comparing minimum free energy structures?
  - *Computationally derived minimum free energy secondary structure are seldom precise*
- Comparing minimum free energy values?
  - *Structural changes might not be reflected in a significant difference of free energy*
- Our approach – comparing Boltzmann distributions

Looking at the differences between structures from the perspective of Boltzmann ensemble

Sampled ensembles of 5'UTR FTL gene (MDS Scaling)

wild type *G32C mutant*

Each circle represents an RNA secondary structure and the size of the circle is proportional to the probability of the structure in the corresponding ensemble.



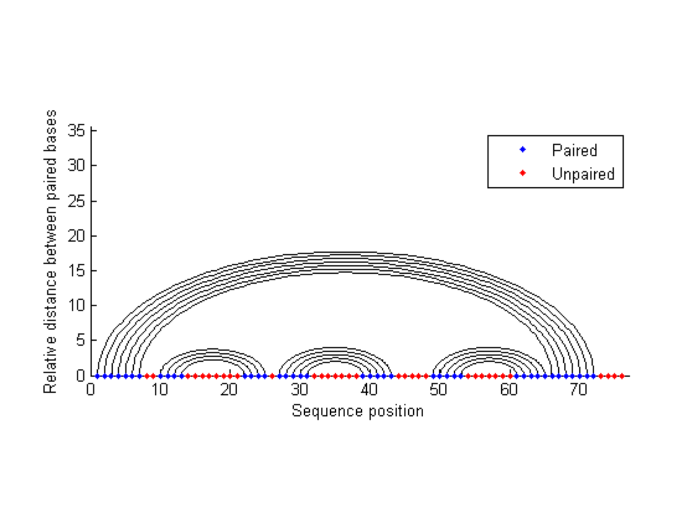
Salari et al. *RECOMB 2012*
Salari et al. *NAR 2012*

Relative Entropy

Kullback–Leibler divergence

$$D_{KL}(wt||mu) = \sum_{s \in S} \mathbb{P}(s|wt) \log(\frac{\mathbb{P}(s|wt)}{\mathbb{P}(s|mu)})$$

Summation over all secondary structures
a direct enumeration computationally intractable!!!



Nesting properties of RNA secondary structure and additivity of energy terms allows dynamic programming strategies

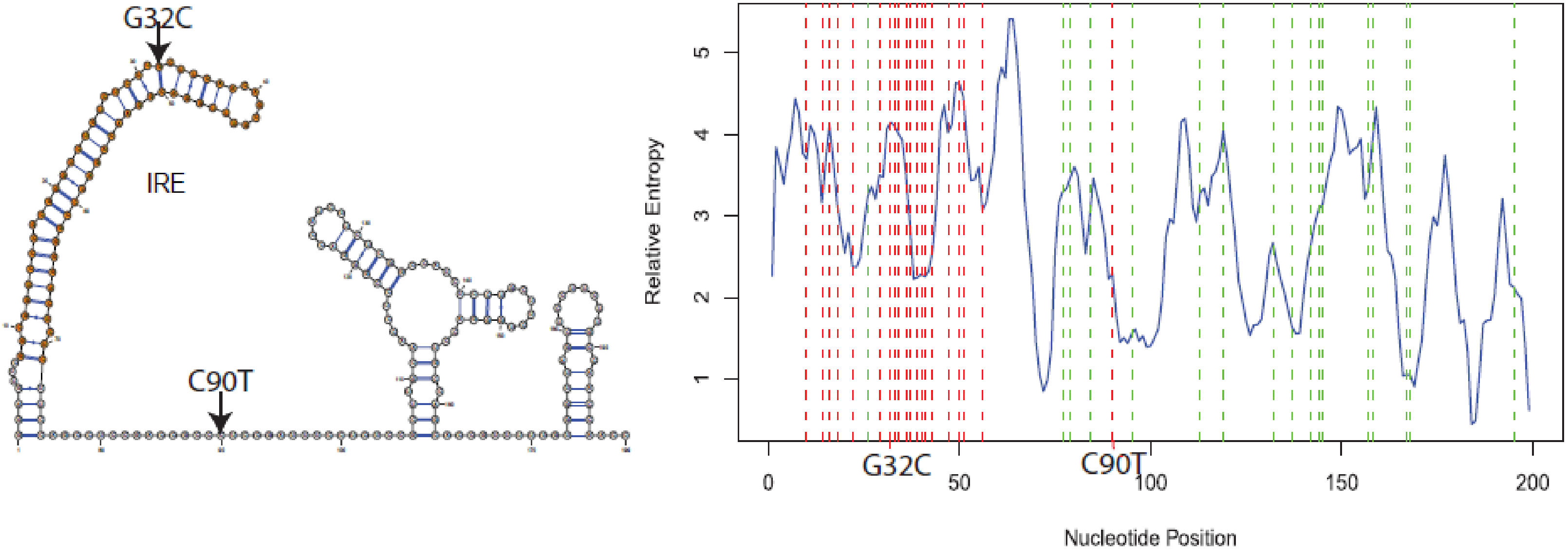
Salari et al. *RECOMB 2012*
Salari et al. *NAR 2012*

$$H_{i,j} = H_{i+1,j} + \sum_{i<k<j} H^b_{i,k} Q_{k+1,j} + \sum_{i<k<j} Q^b_{i,k} H_{k+1,j}.$$

$$\begin{aligned}
H^b_{i,j} = {} & e^{-G^H_{wt}(i,j)/RT}[G^H_{wt}(i,j) - G^H_{mu}(i,j)]/RT \\
& + \sum_{i<k<l<j} Q^b_{k,l} e^{-G^I_{wt}(i,k,l,j)/RT} \\
& [G^I_{wt}(i,k,l,j) - G^I_{mu}(i,k,l,j)]/RT \\
& + \sum_{i<k<l<j} H^b_{k,l} e^{-G^I(i,k,l,j)/RT} \\
& + \sum_{i<k<l<j} e^{-(G^I_{wt}(i,k,l,j)+G^H_{wt}(k,l))/RT} \\
& [G^I_{wt}(i,k,l,j) + G^H_{wt}(k,l) - G^I_{mu}(i,k,l,j) - G^H_{mu}(k,l)]/RT \\
& + \sum_{\substack{i<k<u< \\ v<l<j}} Q^b_{u,v} e^{-(G^I_{wt}(i,k,l,j)+G^I_{wt}(k,u,v,l))/RT} \\
& [G^I_{wt}(i,k,l,j) + G^I_{wt}(k,u,v,l) \\
& - G^I_{mu}(i,k,l,j) - G^I_{mu}(k,u,v,l)]/RT \\
& + \sum_{\substack{i<k<u< \\ v<l<j}} Q^b_{u,v} Q^m_{v+1,l-1} e^{-(G^M(u-k-1,1)+G^I_{wt}(i,k,l,j))/RT} \\
& [G^I_{wt}(i,k,l,j) - G^I_{mu}(i,k,l,j)]/RT \\
& + \sum_{i<k<l<j} H^b_{k,l} Q^m_{l+1,j-1} \;\; e^{-G^M(k-i-1,1)/RT} \\
& + \sum_{i<k<l<j} Q^b_{k,l} H^m_{l+1,j-1} \;\; e^{-G^M(k-i-1,1)/RT}.
\end{aligned}$$

$$\begin{aligned}
H^m_{i,j} = {} & \sum_{i \le k<l \le j} H^b_{k,l} \;\; e^{-(\alpha_2(k-i+j-l)+\alpha_3)/RT} \\
& + \sum_{i \le k<l<j} H^b_{k,l} Q^m_{l+1,j} \;\; e^{-(\alpha_2(k-i)+\alpha_3)/RT} \\
& + \sum_{i \le k<l<j} Q^b_{k,l} H^m_{l+1,j} \;\; e^{-(\alpha_2(k-i)+\alpha_3)/RT}.
\end{aligned}$$

Stability profile



(a) Predicted Secondary Structure

(b) Stability Profile

Red – know disease causing mutations
Green – common SNPs

Disease associated mutations that induce changes in RNA stucture

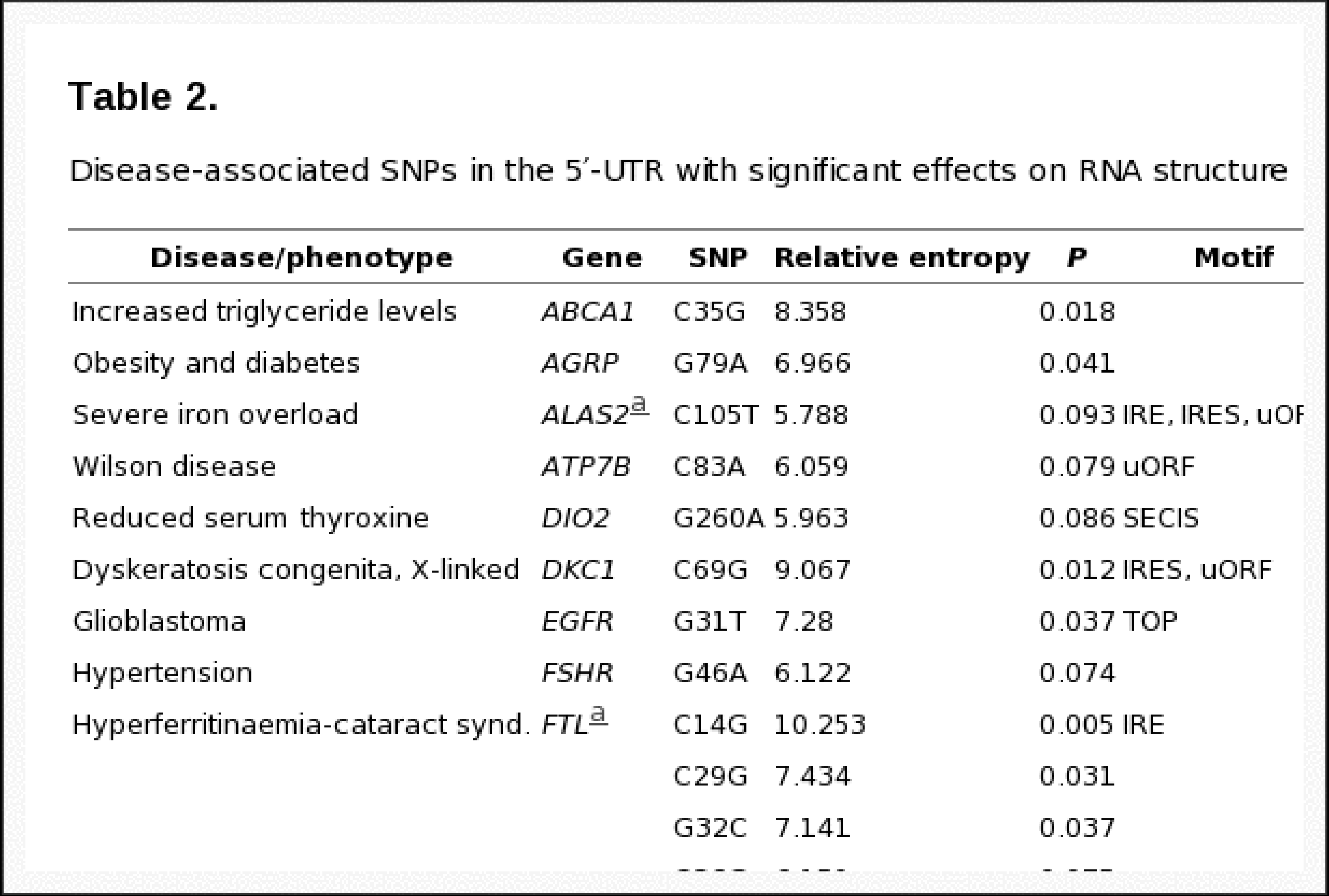


Table 2.

Disease-associated SNPs in the 5′-UTR with significant effects on RNA structure

Disease/phenotype	Gene	SNP	Relative entropy	*P*	Motif
Increased triglyceride levels	*ABCA1*	C35G	8.358	0.018	
Obesity and diabetes	*AGRP*	G79A	6.966	0.041	
Severe iron overload	*ALAS2*[a]	C105T	5.788	0.093	IRE, IRES, uOF
Wilson disease	*ATP7B*	C83A	6.059	0.079	uORF
Reduced serum thyroxine	*DIO2*	G260A	5.963	0.086	SECIS
Dyskeratosis congenita, X-linked	*DKC1*	C69G	9.067	0.012	IRES, uORF
Glioblastoma	*EGFR*	G31T	7.28	0.037	TOP
Hypertension	*FSHR*	G46A	6.122	0.074	
Hyperferritinaemia-cataract synd.	*FTL*[a]	C14G	10.253	0.005	IRE
		C29G	7.434	0.031	
		G32C	7.141	0.037	

Natural polymorphism has smaller impact on mRNA structure that randomly inserted mutations

SNP Class	P-value
CDs	4e-4
5’UTR	7e-3
3’UTR	1e-6

Part I summary

- We have developed a method to subtle structural changes introduced by SNV

- Our method can help to identify dieses causing mutations that can act by structure changes

- Can be used to study impact in structure on the evolution of protein coding sequences

RNA/ssDNA aptamers

Aptamers -small nucleic acid molecules that bind to a target molecule (or a cell)

- Potential to inhibit the biological function of the molecule
- Can be used to differentiate molecules or cell type (molecular testing)
- Antibody replacement

Identification of aptamers with the SELEX protocol
Systematic Evolution of Ligands by Exponential Enrichment

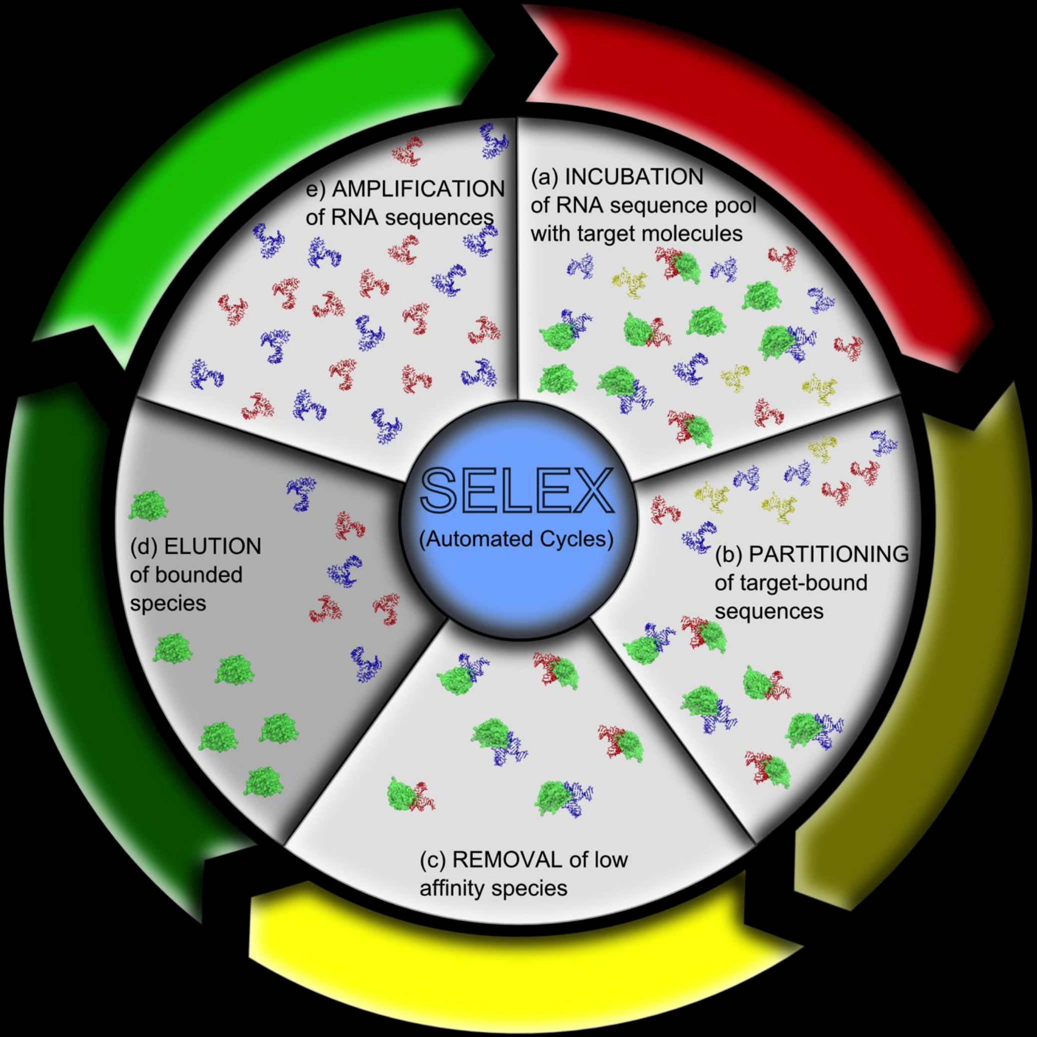
SELEX Protocol (~1990) :

Given

- **(**a random) pool of RNA or ssDNA molecules
- binding target

Goal

- select the binders to the target



SELEX
(Automated Cycles)
(a) INCUBATION of RNA sequence pool with target molecules
(b) PARTITIONING of target-bound sequences
(c) REMOVAL of low affinity species
(d) ELUTION of bounded species
e) AMPLIFICATION of RNA sequences

Identification of Aptamers with the SELEX protocol
Systematic Evolution of Ligands by Exponential Enrichment

Traditionally – a black box procedure

Computational approaches allow for a more insightful application of this technology

- Identification of binding motifs that account for sequence –structure properties - Aptamotif approach

Hoinka et al. *ISMB* 2012, *Bioinformatics* 2012

- Computational methods for the analysis of the results of HT-SELEX – HTAptamotif

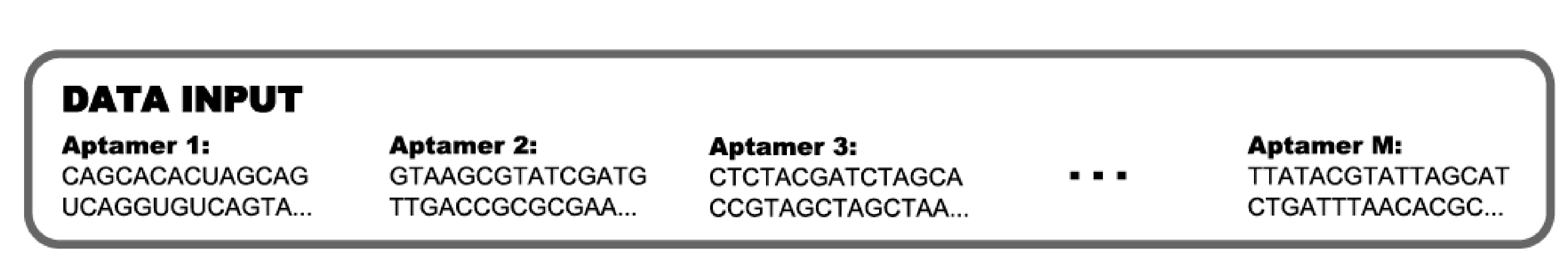
Current work

Aptamotif – Identification of sequence – structure binding motives in traditional SELEX experiments

Underlying assumptions of the approach:

- Binding motifs are in loop regions
- Binding motifs do not need to be contiguous
- A biding motifs is restricted to one loop rather than distributed over several loops
- The binding conformation does not have to correspond to the minimum free energy structure

Hoinka et al. *ISMB* 2012, *Bioinformatics* 2012



DATA INPUT

Aptamer 1:
CAGCACACUAGCAG
UCAGGUGUCAGTA...

Aptamer 2:
GTAAGCGTATCGATG
TTGACCGCGCGAA...

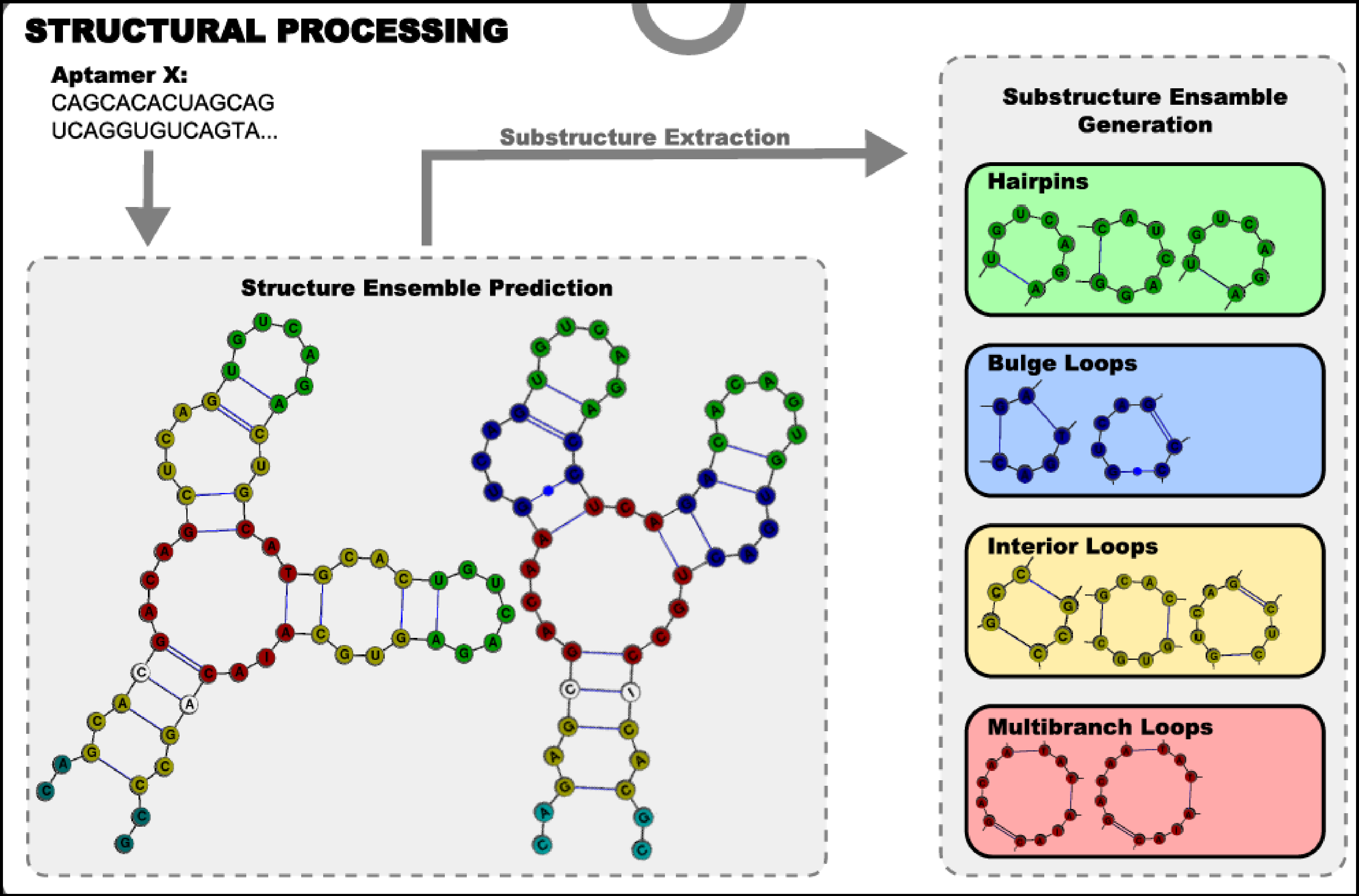
Aptamer 3:
CTCTACGATCTAGCA
CCGTAGCTAGCTAA...

. . .

Aptamer M:
TTATACGTATTAGCAT
CTGATTTAACACGC...

For each aptamer generate optimal and suboptimal secondary structures

Decompose all optimal and suboptimal structures into loops



STRUCTURAL PROCESSING

Aptamer X:
CAGCACACUAGCAG
UCAGGUGUCAGTA...

Structure Ensemble Prediction

Substructure Extraction

Substructure Ensamble Generation

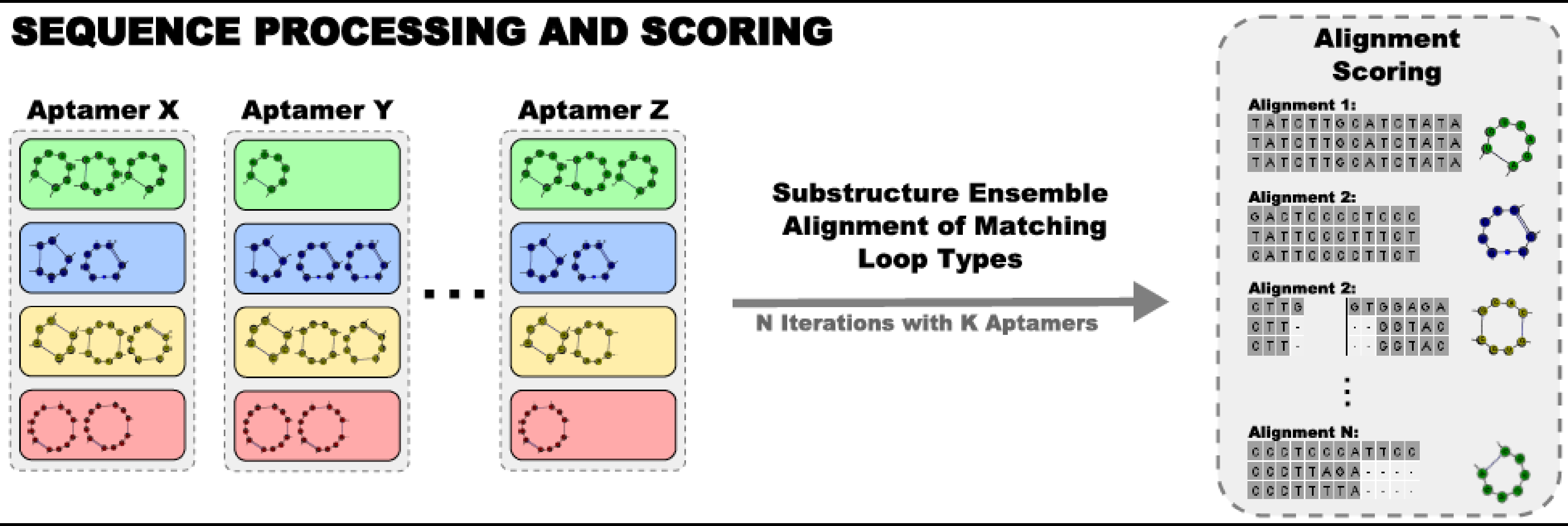
Hairpins

Bulge Loops

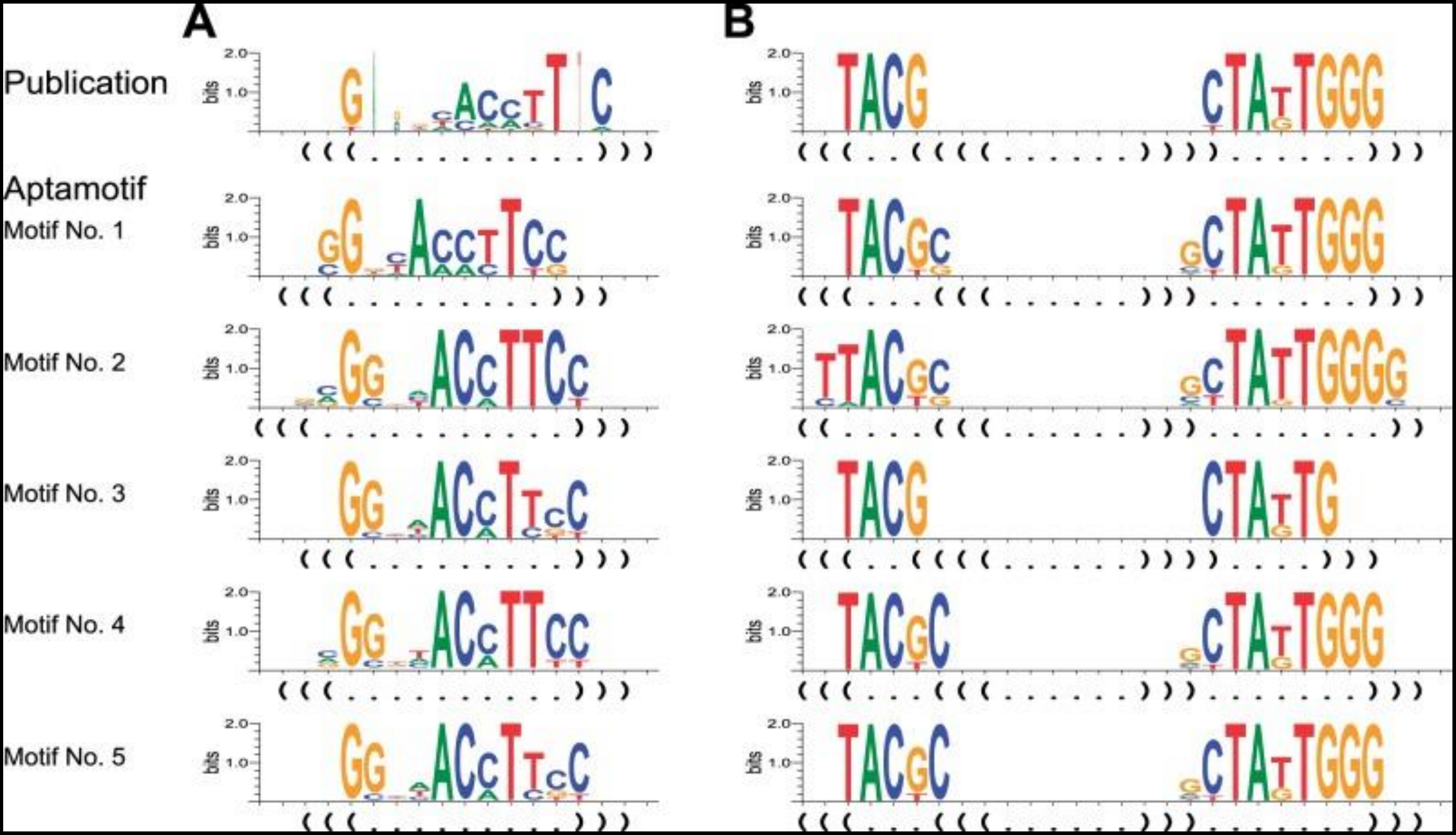
Interior Loops

Multibranch Loops

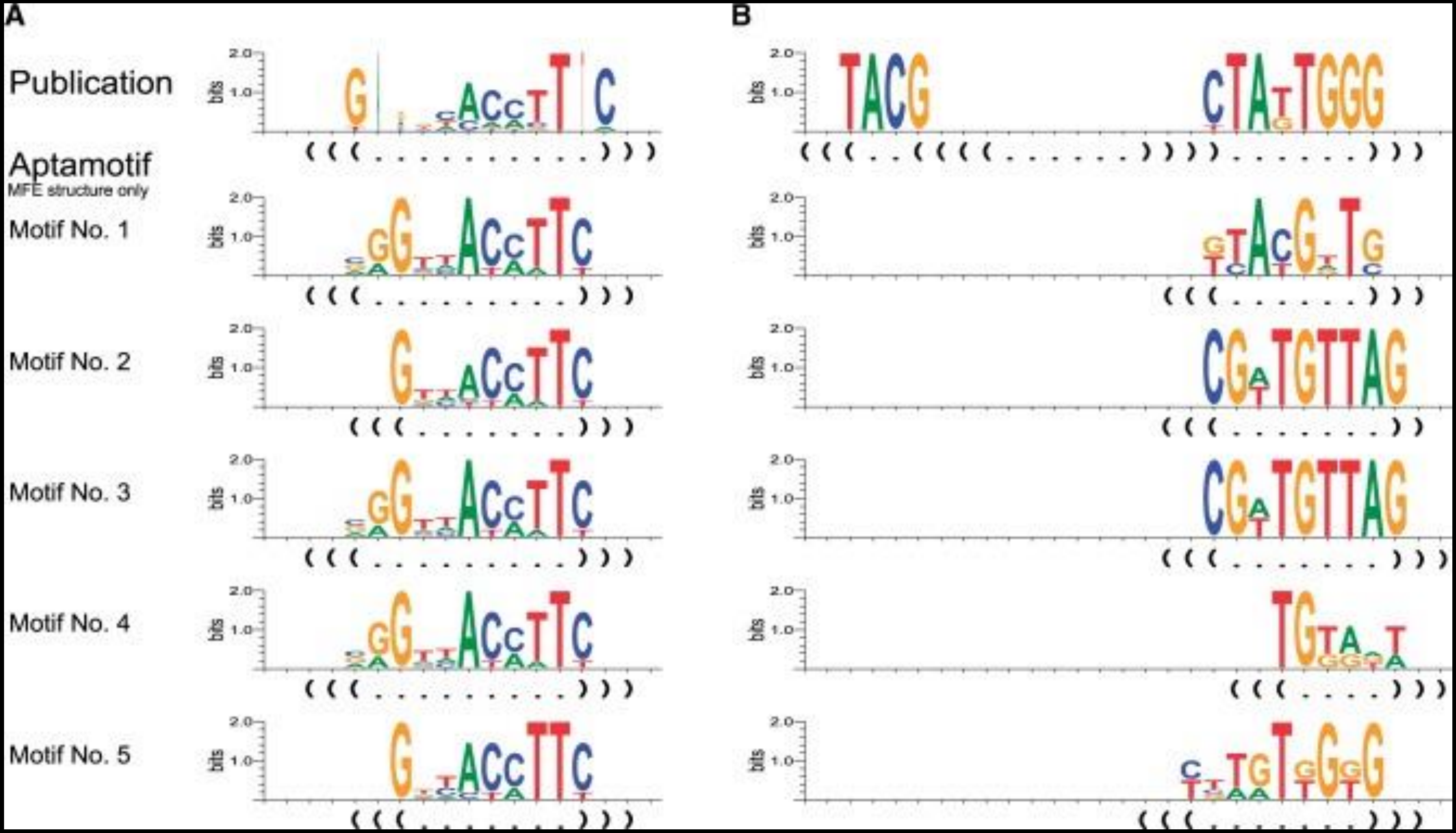
For each loop type find enriched sequence motifs



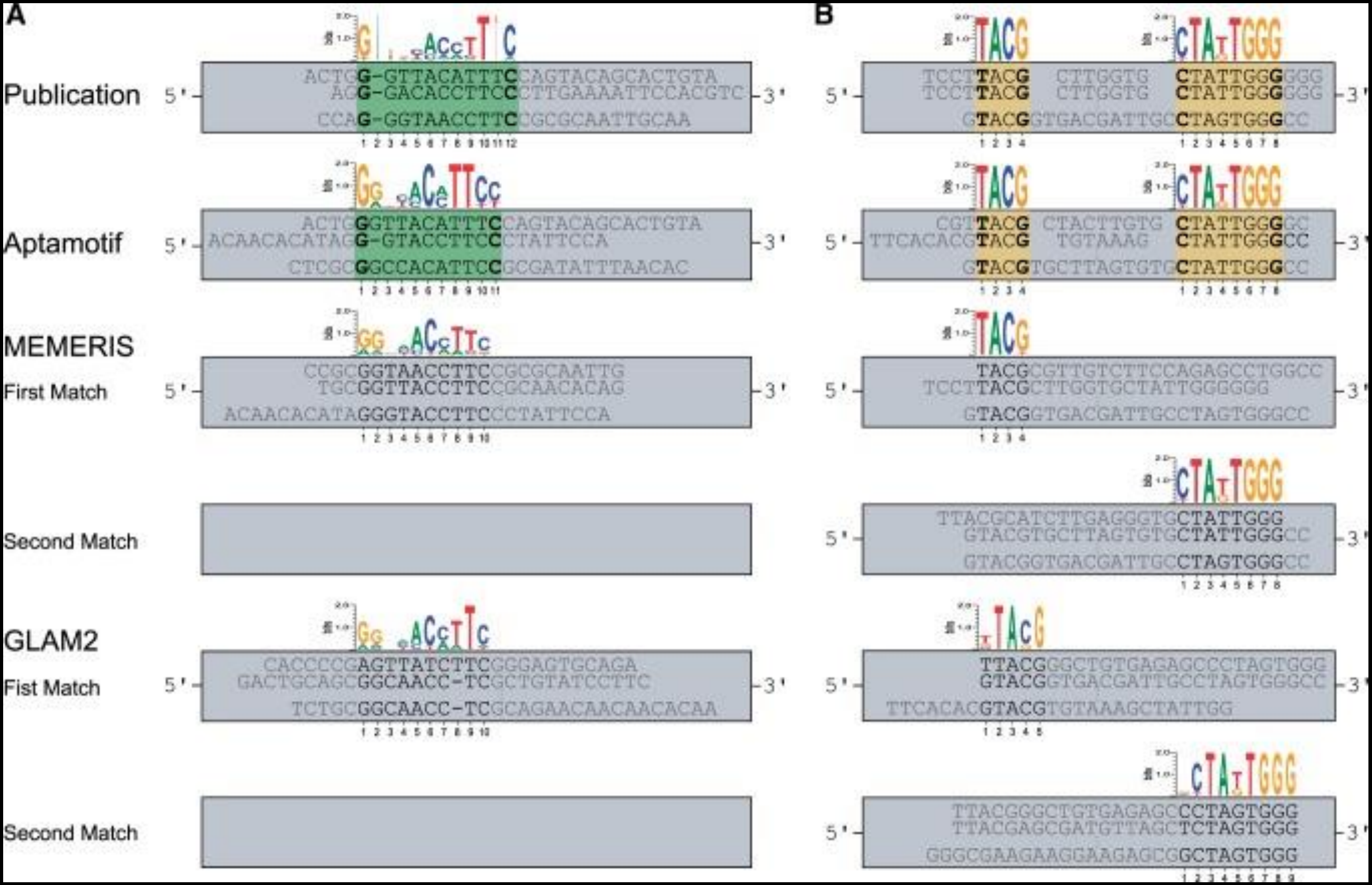
First five top scoring motifs for the datasets of (**A**) Dobbelstein and Shenk (1995) and (**B**) Lozupone *et al.* (2003) identified by *Aptamotif*



Using only minimum free energy structure doesn't work



Comparing with motif finding approaches


A
B
Publication
Aptamotif
MEMERIS
First Match
Second Match
GLAM2
Fist Match
Second Match
5'
3'
ACTGG-GTTACATTTCCAGTACAGCACTGTA
ACG-GACACCTTCCCTTGAAAATTCCACGTC
CCAG-GGTAACCTTCCGCGCAATTGCAA
ACTGGGTTACATTTCCAGTACAGCACTGTA
ACAACACATAGG-GTACCTTCCCTATTCCA
CTCGCGGCCACATTCCGCGATATTTAACAC
CCGCGGTAACCTTCCGCGCAATTG
TGCGGTTACCTTCCGCAACACAG
ACAACACATAGGGTACCTTCCCTATTCCA
CACCCCGAGTTATCTTCGGGAGTGCAGA
GACTGCAGCGGCAACC-TCGCTGTATCCTTC
TCTGCGGCAACC-TCGCAGAACAACAACACAA
TCCTTACG CTTGGTG CTATTGGGGGG
TCCTTACG CTTGGTG CTATTGGGGGG
GTACGGTGACGATTGCCTAGTGGGCC
CGTTACG CTACTTGTG CTATTGGGGC
TTCACACGTACG TGTAAAG CTATTGGGCC
GTACGTGCTTAGTGTGCTATTGGGCC
TACGCGTTGTCTTCCAGAGCCTGGCC
TCCTTACGCTTGGTGCTATTGGGGGG
GTACGGTGACGATTGCCTAGTGGGCC
TTACGCATCTTGAGGGTGCTATTGGG
GTACGTGCTTAGTGTGCTATTGGGCC
GTACGGTGACGATTGCCTAGTGGGCC
TTACGGGCTGTGAGAGCCCTAGTGGG
GTACGGTGACGATTGCCTAGTGGGCC
TTCACACGTACGTGTAAAGCTATTGG
TTACGGGCTGTGAGAGCCCTAGTGGG
TTACGAGCGATGTTAGCTCTAGTGGG
GGGCGAAGAAGGAAGAGCGGCTAGTGGG

Aptamotif summary

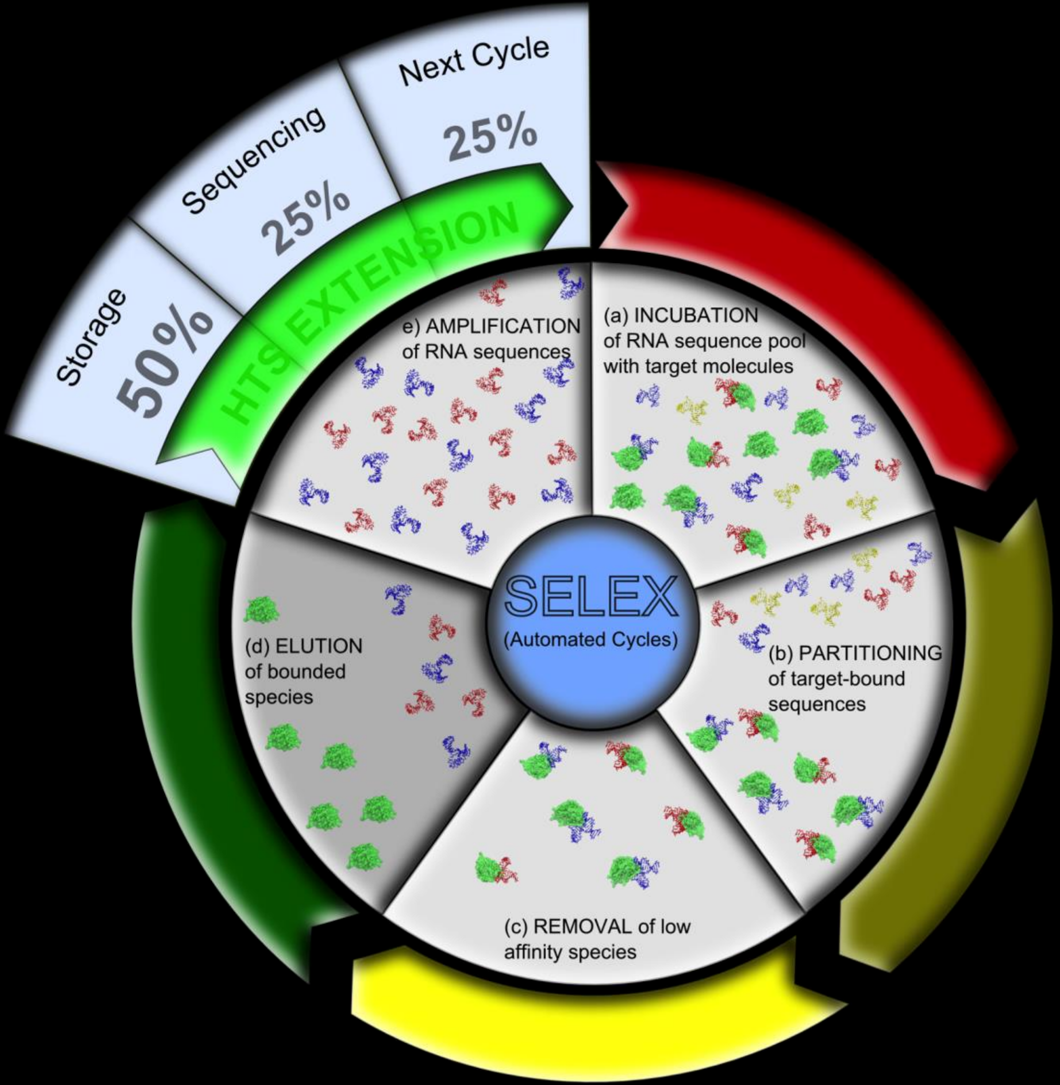
- Importance of secondary structure
- Importance of sampling of suboptimal structures

Things of potential importance we haven't consider due to the lack of supporting data

- Sequence specificity of non-loop regions
- Combinatorial effect of many loops

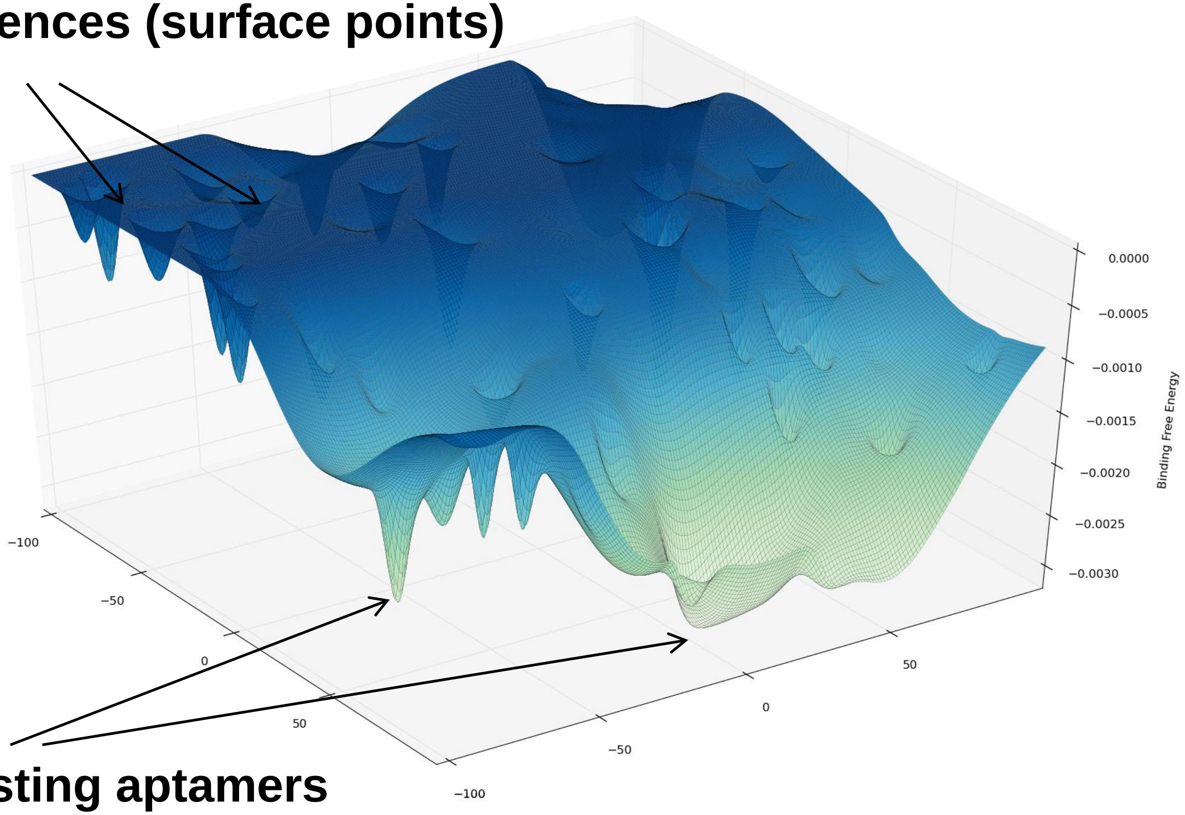
HT-SELEX– a new powerful variant of the SELEX experiment

Next-gen sequencing of a *samples* of   intermediate selection pools


Storage
50%
Sequencing
25%
Next Cycle
25%
HTS EXTENSION
e) AMPLIFICATION of RNA sequences
(a) INCUBATION of RNA sequence pool with target molecules
(b) PARTITIONING of target-bound sequences
(c) REMOVAL of low affinity species
(d) ELUTION of bounded species
SELEX
(Automated Cycles)

Hoinka et al, current work

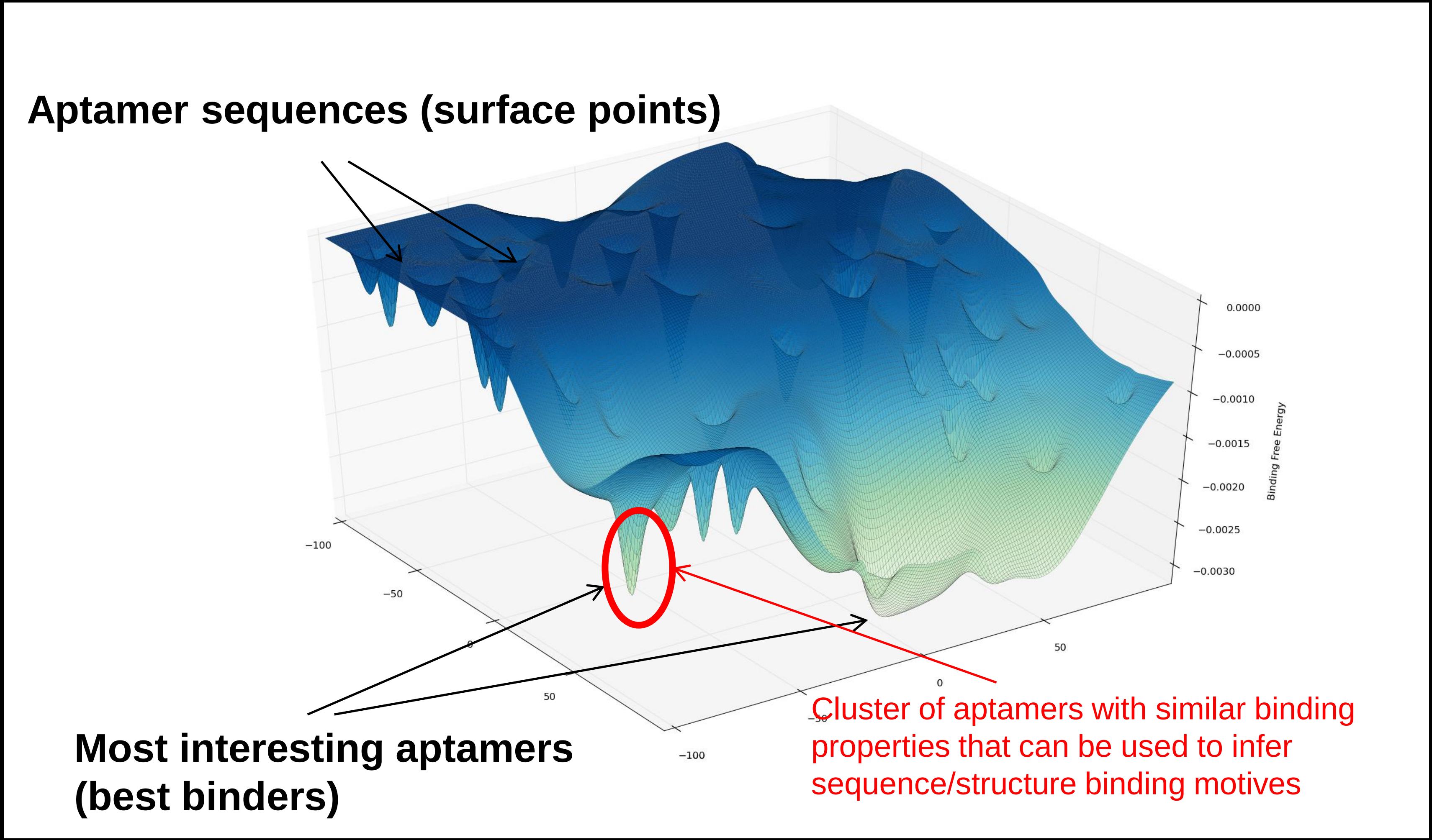
Potential opportunity – delineating binding energy landscape



Aptamer sequences (surface points)

Most interesting aptamers (best binders)

Potential opportunity – delineating binding energy landscape

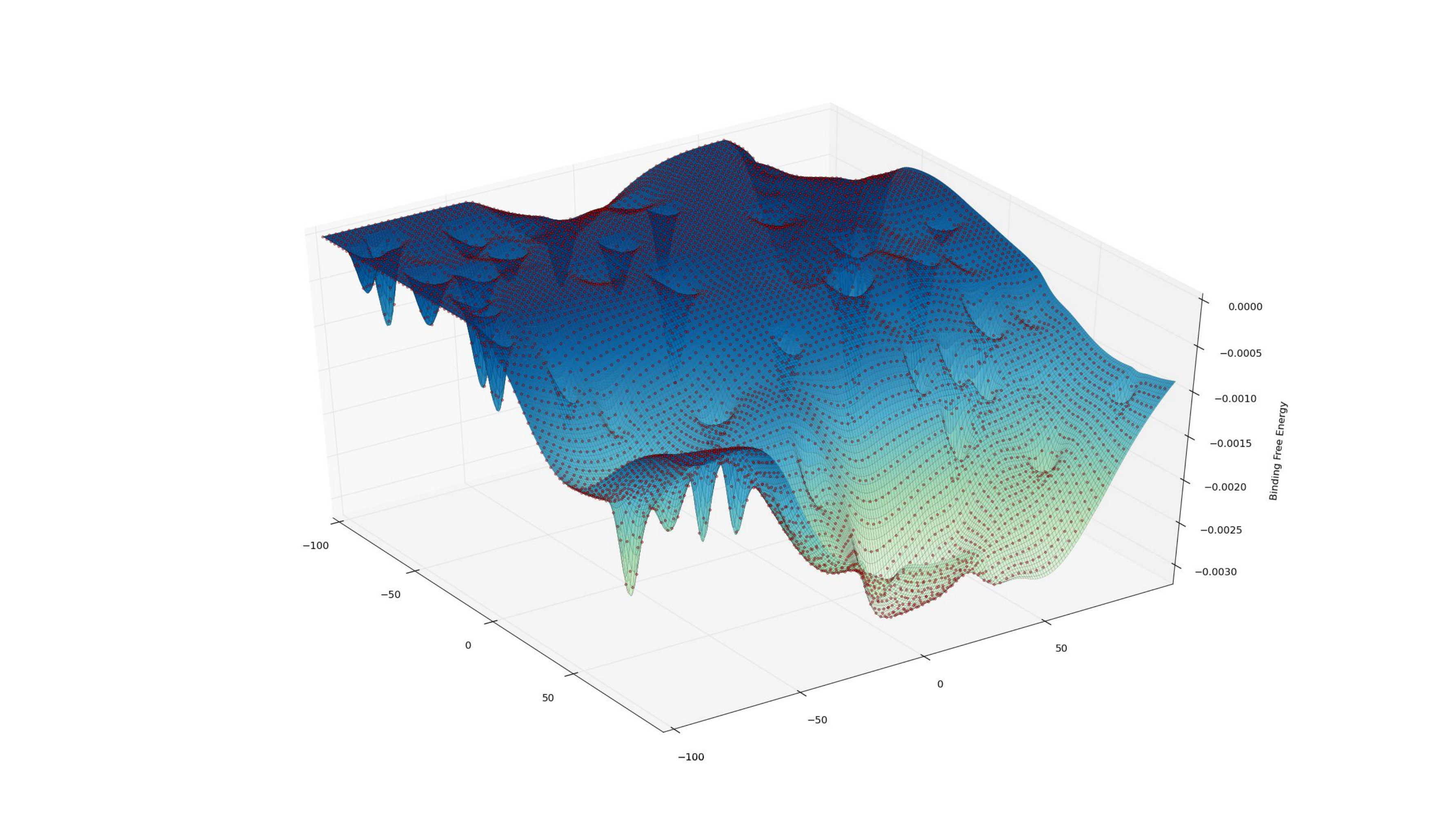


Aptamer sequences (surface points)

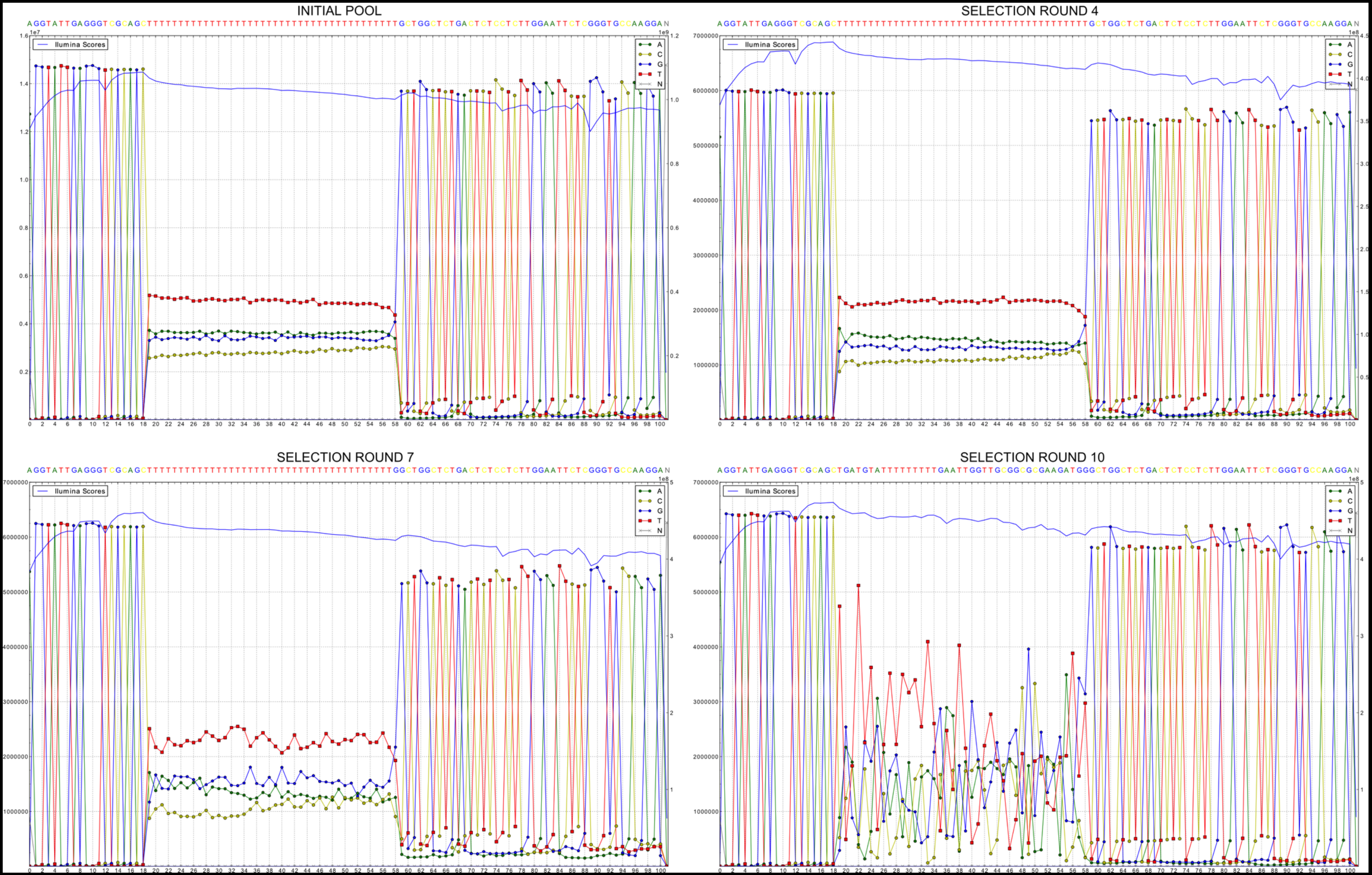
Most interesting aptamers (best binders)

Cluster of aptamers with similar binding properties that can be used to infer sequence/structure binding motives

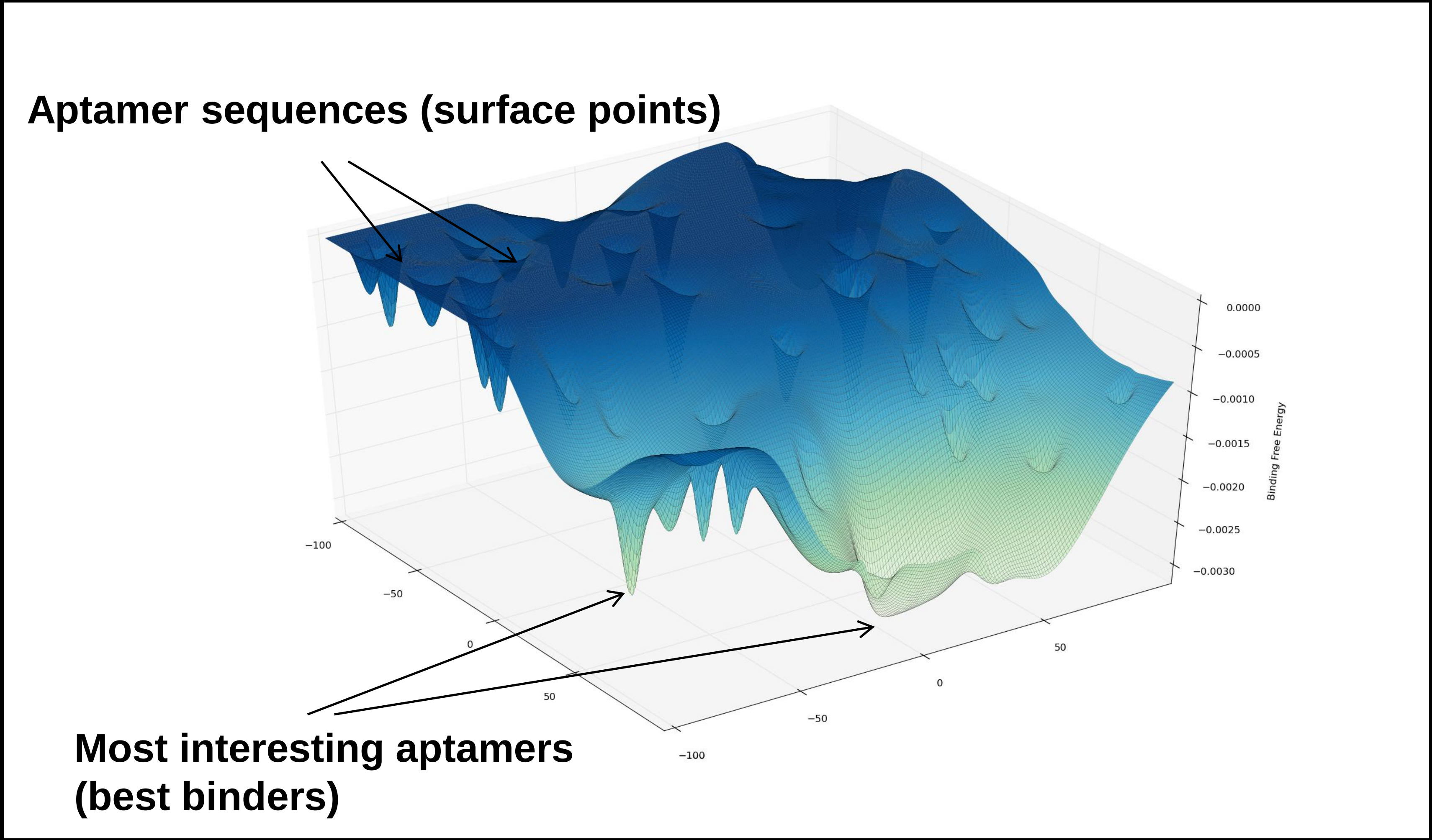
Wishful thinking #1: we start with a uniform sampling of the aptamer space



Not true


INITIAL POOL
AGGTATTGAGGGTCGCAGCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGCTGGCTCTGACTCTCCTCTTGGAATTCTCGGGTGCCAAGGAN
SELECTION ROUND 4
AGGTATTGAGGGTCGCAGCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGCTGGCTCTGACTCTCCTCTTGGAATTCTCGGGTGCCAAGGAN
SELECTION ROUND 7
AGGTATTGAGGGTCGCAGCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGGCTGGCTCTGACTCTCCTCTTGGAATTCTCGGGTGCCAAGGAN
SELECTION ROUND 10
AGGTATTGAGGGTCGCAGCTGATGTATTTTTTTTTGAATTGGTTGCGGCGCGAAGATGGGCTGGCTCTGACTCTCCTCTTGGAATTCTCGGGTGCCAAGGAN
Ilumina Scores
A
C
G
T
N
1e7
1e9
1e8
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84 86 88 90 92 94 96 98 100
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
7000000
6000000
5000000
4000000
3000000
2000000
1000000
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
5
4
3
2
1

Wishful thinking#2: most abundant aptamers that at the end of the selection process are the best binders



Aptamer sequences (surface points)

Most interesting aptamers (best binders)

Only partially true

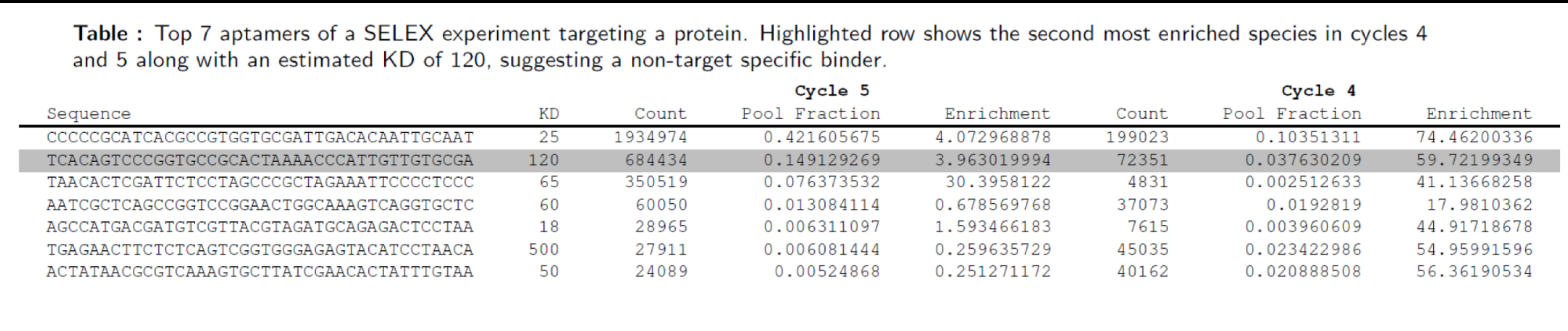


Table : Top 7 aptamers of a SELEX experiment targeting a protein. Highlighted row shows the second most enriched species in cycles 4 and 5 along with an estimated KD of 120, suggesting a non-target specific binder.

			Cycle 5			Cycle 4	
Sequence	KD	Count	Pool Fraction	Enrichment	Count	Pool Fraction	Enrichment
CCCCCGCATCACGCCGTGGTGCGATTGACACAATTGCAAT	25	1934974	0.421605675	4.072968878	199023	0.10351311	74.46200336
TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	120	684434	0.149129269	3.963019994	72351	0.037630209	59.72199349
TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	65	350519	0.076373532	30.3958122	4831	0.002512633	41.13668258
AATCGCTCAGCCGGTCCGGAACTGGCAAAGTCAGGTGCTC	60	60050	0.013084114	0.678569768	37073	0.0192819	17.9810362
AGCCATGACGATGTCGTTACGTAGATGCAGAGACTCCTAA	18	28965	0.006311097	1.593466183	7615	0.003960609	44.91718678
TGAGAACTTCTCTCAGTCGGTGGGAGAGTACATCCTAACA	500	27911	0.006081444	0.259635729	45035	0.023422986	54.95991596
ACTATAACGCGTCAAAGTGCTTATCGAACACTATTTGTAA	50	24089	0.00524868	0.251271172	40162	0.020888508	56.36190534

Possible causes:

- Biased sampling
- Amplification bias
- Non specific binding
- Discrepancy between sequenced sample and amplified sample
- Aptamer mutants

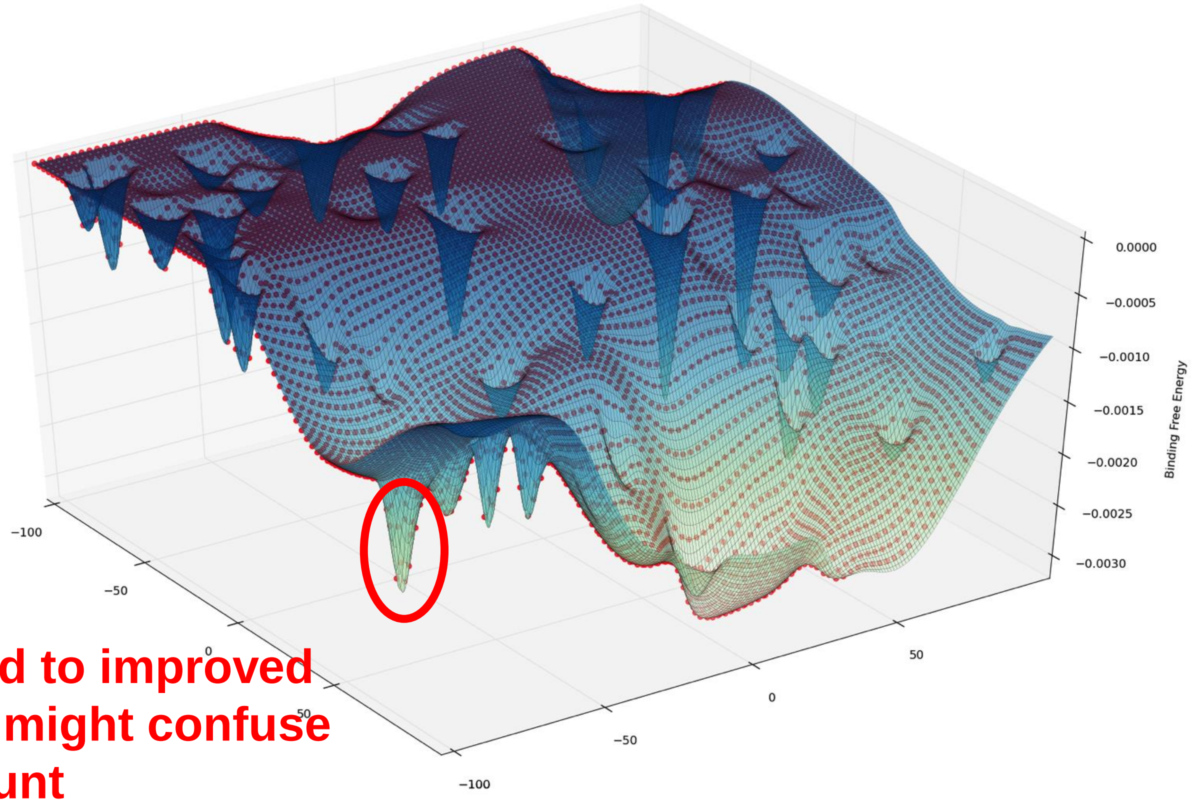
Promising predictors:

- Cycle – to - cycle enrichment (excluding very early/late cycles)

Interesting opportunity:

- Tracing enrichment of Aptamer mutants

Tracing Aptamers mutations introduced by Pol II



Mutations lead to improved sampling but might confuse frequency count

HT-Aptamotif – toolbox to the analysis of HT-SELEX data and the identification sequence-structure motifs (under development)

People contributing data and experimental expertise

Eli Gilboa and to Alex Berezhnoy University of Miami
Zuben Sauna, FDA
Scott D. Rose and Mark Behlke, Integrated DNA Technologies
Rebecca Whelan, Oberlin College

Computational analysis

- Quality control (and some error correction tools)
- Sequence based clustering* (will put the mutants with parent sequence)

  *clustering huge aptamer pool is computationally challenging
- Cycle-to-cycle cluster enrichment analysis
- Identification of sequence-structure motifs

current work

SUMMARY

Ensemble approach was fundamental to

- Measuring impact of a SNV on RNA structure
- delineating sequence structure motifs

The presentation utilized data from

Eli Gilboa to Alex Berezhnoy University of Miami
Rebecca Whelan, Oberlin College

Acknowledgments

Przytycka's group

DongYeon Cho
- *Prob. Cancer Model*
- *CNV in fly*

Phuong Dao
- Gene regulation

Xiangjun Du
- *Non B-DNA*

Jan Hoinka
- *Aptamers*

Yoo-Ah Kim
- *Cancer networks*
- *Gene regulation*

Damian Wojtowicz
- *Non-B-DNA, Promoter Structure*
- *Expression noise*

Former group member

Raheleh Salari (Stanford University)
RNA SNP

Collaborators

(for the discussed topics)

Eli Bilboa & Alex Berezhnoy U. Miami

Michael Gottesman, NCI

Chava Kimchy-Sarfaty, FDA

Zuben Sauna, FDA

Scott D. Rose & Mark Behlke
Integrated DNA Technologies

Rebecca Whelam Oberlin College